DESDE 2023
https://revista.ccaitese.com

RICT Revista de Investigación Científica, Tecnológica e Innovación

Publicación Semestral RICT Vol. 1 No. 2 (2023) p. 32-37

ISSN: 2992-7315

Regresión logística v/s Arboles de decisión en el riesgo crediticio.
Logistic regression v/s decision trees in credit risk.

Luisa-Jaquelin. González-Segoviano[a]

[a] *División de ingeniería en sistemas computacionales, Tecnológico de estudios superiores de Ecatepec, 55210, México.*

Resumen

El riesgo crediticio tiene que ver con la situación en la que un prestatario no cumpla con sus obligaciones de pago. La regresión lineal es una técnica estadística que permite modelar la relación entre una variable dependiente (como el riesgo crediticio) y una o más variables independientes (como los factores que afectan al prestatario). En el contexto del riesgo crediticio, la regresión lineal puede utilizarse para analizar y predecir el nivel de riesgo asociado con un determinado prestatario. Se recopilan datos relevantes sobre el prestatario, como su historial crediticio, ingresos, nivel de endeudamiento, entre otros, y se utiliza la regresión lineal para identificar patrones y relaciones entre estas variables. Otra herramienta son los árboles de decisión que son una técnica de modelado utilizada en el análisis de riesgo crediticio para tomar decisiones basadas en múltiples variables. Permiten evaluar patrones y relaciones entre las variables y ayudan a tomar decisiones para el otorgamiento de crédito. Los árboles de decisión son interpretables y proporcionan una comprensión clara del razonamiento detrás de las decisiones tomadas. El objetivo de este estudio fue comparar el rendimiento de dos algoritmos de aprendizaje supervisado, el Árbol de Decisión y la Regresión Logística, en la predicción de riesgo crediticio. Se evaluó la eficiencia que tienen los modelos de regresión logística contra árboles de decisión para poder predecir el riesgo crediticio, siendo regresión logística el que tuvo mayor eficiencia con el 0.93, mientras tanto arboles de decisión tuvo 0.83 de eficiencia a la hora de entrenar los dos modelos con el mismo número de muestras. La evaluación precisa del riesgo crediticio es fundamental para las instituciones financieras al otorgar préstamos y créditos, lo que ayuda a reducir el riesgo de incumplimiento y las pérdidas.

Palabras Clave: árboles de decisión, riesgo, crédito, regresión lineal, predicción

Abstract

Credit risk has to do with the situation in which a borrower does not meet its payment obligations. Linear regression is a statistical technique that allows you to model the relationship between a dependent variable (such as credit risk) and one or more independent variables (such as factors that affect the borrower). In the context of credit risk, linear regression can be used to analyze and predict the level of risk associated with a given borrower. Relevant data about the borrower is collected, such as their credit history, income, level of indebtedness, among others, and linear regression is used to identify patterns and relationships between these variables. Another tool is decision trees, which are a modeling technique used in credit risk analysis to make decisions based on multiple variables. They allow evaluating patterns and relationships between variables and help make decisions for granting credit. Decision trees are interpretable and provide a clear understanding of the reasoning behind the decisions made. The objective of this study was to compare the performance of two supervised learning algorithms, Decision Tree and Logistic Regression, in predicting credit risk. The efficiency of the logistic regression models against decision trees was evaluated to predict credit risk, with logistic regression being the one that had the highest efficiency with 0.93, while decision trees had 0.83 efficiency when training the two models with the same number of samples. Accurate credit risk assessment is essential for financial institutions when making loans and credit, helping to reduce the risk of default and losses.

Keywords: decision tree, risk, credit, linear regression, prediction.

*Autor para la correspondencia: 202210961@tese.edu.mx
Correo electrónico: 202210961@tese.edu.mx (Luisa-Jaquelin González-Segoviano).

Historial del manuscrito: recibido el 02/08/2023, última versión-revisada recibida el 08/09/2023, aceptado el 09/09/2023, en línea (postprint) desde el 09/09/2023, publicado el 30/09/2023. **DOI:** En trámite

1. Introducción

El riesgo crediticio es una preocupación fundamental en el ámbito financiero, especialmente para las instituciones que otorgan préstamos y créditos. Se refiere a la probabilidad de que un prestatario no cumpla con sus obligaciones de pago, lo que puede generar pérdidas financieras para el prestamista. Por lo tanto, es esencial comprender y evaluar adecuadamente el riesgo crediticio antes de otorgar crédito a un individuo, empresa u otra entidad (Hernández, 2004).

La introducción del riesgo crediticio implica reconocer que toda transacción crediticia conlleva ciertos niveles de incertidumbre. Esto se debe a que existen diversos factores que pueden afectar la capacidad y disposición del prestatario para cumplir con sus obligaciones.

Algunos de estos factores incluyen la calidad crediticia del prestatario, su historial de pagos, su capacidad de generar ingresos suficientes para cubrir las obligaciones de deuda, la situación económica general y otros riesgos específicos asociados con el sector o la industria en la que opera el prestatario (Pérez, 2017).

Los bancos son intermediarios financieros que recaudan fondos públicos e invierten esos fondos en préstamos e inversiones. Diversas actividades, como los depósitos y colocaciones bancarias, implican riesgos financieros. Es decir, las pérdidas económicas que pueden derivarse del propio negocio bancario. En este sentido, si un banco otorga un préstamo, es posible que el prestatario no lo devuelva. Esto se conoce como riesgo de crédito (Díaz & Del Valle Guerra, 2017).

Una sociedad financiera comercial es una sociedad cuyo objeto principal es la captación de recursos a corto plazo con el fin de realizar operaciones activas de préstamo para facilitar la comercialización de bienes y servicios. Para las empresas que brindan servicios financieros de crédito, es fundamental contar con liquidez financiera para cumplir con todas las obligaciones asumidas y un flujo de efectivo estable para continuar brindando los servicios. Además, este tipo de empresas también buscan mitigar el riesgo de crédito por la naturaleza de su negocio (Diego Borrero-Tigreros, 2020).

El otorgamiento de crédito es parte integral del apalancamiento de los fondos de los empleados y otras instituciones financieras además de los beneficios que reciben los empleados de estas empresas para su crecimiento personal y el crecimiento de la sociedad en su conjunto (Vergara, 2022).

Fundamentalmente, el negocio de una institución financiera implica riesgo y todas sus operaciones están sujetas a incertidumbres implícitas y explícitas. El negocio de la empresa tiene múltiples facetas y, por tanto, está expuesto a diferentes tipos de contingencias que es necesario identificar, medir y controlar para que sirvan de base para establecer una estrategia de marketing, en particular un precio que proporcione una ecuación favorable entre las contingencias asumidas y la diferenciación lograda y el resultado final de la empresa (Hernández, 2004).

Los clientes incumplen cuando alcanzan un nivel de incumplimiento en el que el banco asume una pérdida de capital.

Por lo tanto, la pérdida esperada se calcula como:

$$\boldsymbol{PE = PI * S * E}$$

dónde,

PI: Probabilidad de falla en un período de tiempo específico.

S: LGD: El porcentaje de la cantidad que la empresa se arriesga a perder si el deudor no cumple con sus obligaciones.

E: Exposición: La cantidad de activos expuestos al riesgo de incumplimiento durante un período de tiempo definido.

La Figura 1 muestra la posibilidad de medir a los clientes en función de su solvencia.

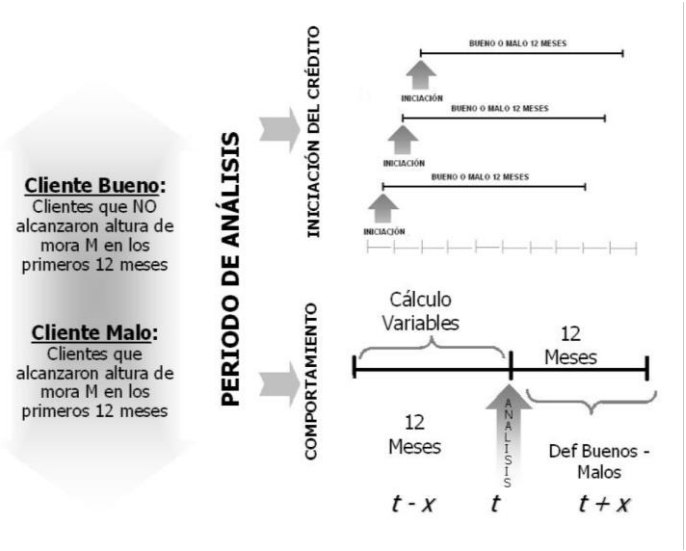

Figura 1 Variables crédito (Hernández, 2004)

Para el análisis se utiliza métodos predictivos, en este artículo se hablará de árboles de decisión y regresión lineal para el análisis de este tema y se hará uso de Python para poder analizar datos y se tenga una comparativa de la mejor técnica.

1.1. Regresión lineal

Desde sus inicios *(MIRANDA, 2013)*, la regresión logística ha sido aceptada como un modelo predictivo de clasificación en estudios epidemiológicos y clínicos durante las últimas décadas. Ahora se usa comúnmente en, entre otros, investigación biomédica, economía, finanzas, criminología, ingeniería, salud pública, política, biología de la vida silvestre y psicología. Las aplicaciones iban desde ganar créditos sin pagar, votar en elecciones políticas, otorgar becas, hasta predecir el cáncer de próstata y el riesgo de infección por VIH.

De acuerdo con Fernández Castaño y Pérez Ramírez (2005), la regresión logística se utiliza cuando se necesita predecir resultados binarios, como spam o no spam, quiebra o no quiebra, y se sabe que hay varios factores que pueden influir en dichos resultados. Dado que el resultado sólo son dos calores 0 a 1, se le llama regresión binomial donde el valor de la variable dependiente es una variable que muestra una probabilidad entre 0 y 1 por lo que al redondearla se obtendrán los valores de 0 o 1. La regresión está basada en la siguiente ecuación:

$$Y_i = \frac{1}{1+exp(-z)} + u_i$$

Donde,

Yi: variable dependiente. Puede tomar valores de 0 o 1.
z: evaluación logística.
u: variable aleatoria normalmente distribuida N(0,s2)

Las variables independientes son fijas dentro de la muestra.
La Regresión Logística Binaria (RLB) es actualmente una herramienta de inferencia que estima coeficientes estadísticos producidos por el algoritmo de Walker-Duncan para obtener estimadores de máxima verosimilitud. En general, la regresión logística es adecuada cuando la respuesta Y es binaria. H. Cuando sólo hay dos respuestas posibles. Este es un buen método para calcular las probabilidades de incumplimiento, pero a pesar de esta calidad estadística, tiene algunas características que limitan su poder predictivo. El comportamiento de los datos es probablemente el más notable, ya que RLB puede dar resultados similares a la regresión lineal *(Pérez, 2017)*.

La regresión logística es en realidad una herramienta eficaz para la clasificación de dos clases y multiclase, es rápida y sencilla. Al utilizar una curva con forma de S en lugar de una línea recta la hace ideal para dividir los datos en grupos *(Microsoft, 2016)*.Whitten y Frank (2005) indican que la regresión logística construye un modelo lineal basado en un variable objetivo. Supongamos primero que sólo hay dos clases.

1.2. Arboles de decisión
Un árbol de decisión es un modelo predictivo formado por reglas binarias (sí/no) que distribuyen las observaciones según atributos y permiten predecir el valor de una variable de respuesta *(Rodrigo, 2020)*.Los métodos basados en árboles se han convertido en referentes en el campo de la predicción porque producen excelentes resultados para una amplia variedad de problemas. Este documento examina cómo se construyen y predicen los árboles de decisión (clasificación y regresión), que son los componentes básicos de modelos predictivos más complejos, como bosques aleatorios y máquinas de aumento de gradiente *(Moreno Villalba, Avila Camacho, & Melendez Ramirez, 2019)*.

Ventajas
• Los árboles son fáciles de interpretar, incluso cuando las relaciones entre los predictores son complejas.
• Los modelos basados en un solo árbol (bosques aleatorios, excepto boosting) se pueden representar incluso cuando el número de predictores supera los tres.
• En teoría, los árboles de decisión se pueden utilizar con predictores numéricos o predictores categóricos sin la necesidad de agregar variables ficticias. Lo cual dependerá del tipo de librería o biblioteca que se utilice.
• Los árboles de decisión son muy útiles para la exploración de datos, permitiéndote identificar las variables más importantes (predictores) de manera rápida y eficiente.
•Seleccionar predictores automáticamente.
•Aplicable a problemas de regresión y clasificación.

Desventajas
•Sensible a los datos de entrenamiento desequilibrados (una clase domina a la otra).
• Cuando se trabaja con predictores continuos, parte de esa información se pierde debido a la clasificación durante la división de nodos.
• No puede extrapolar fuera del rango de predictores observados en los datos de entrenamiento.

2. Materiales y Método

En la Figura 2 se muestra la arquitectura para comparar los árboles de desición contra regressión logística.

El ingreso de datos esta compuesto de las variables de entrada, siendo el ingreso mensual una de las variables más importantes a analizar.Posteriormente se inicia con la etapa de limpieza de datos, donde se eliminaran datos vacios.Es muy importante la etapa de transformación para generar un análisis adecuado y proceder a el entreno con los 2 metodos planteados.

En el resultado es la salida que genera el comportamiento del algoritmo y se analiza la eficiencia de cada uno.

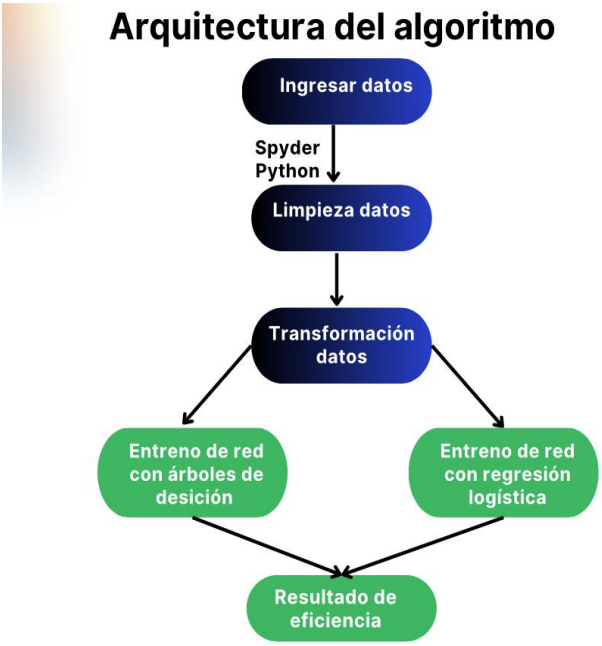

Figura 2 Arquitectura de flujo.

El primer método es regresión logística, con los datos de edad y de tasa de endeudamiento, es necesario importar bibliotecas y leer el archivo que contiene los datos.

En la Figura 3, se muestran las bibliotecas de Python que se requieren para hacer el análisis y a su vez la comparación, todas son muy importantes y sirven para graficar los resultados junto con la eficiencia.

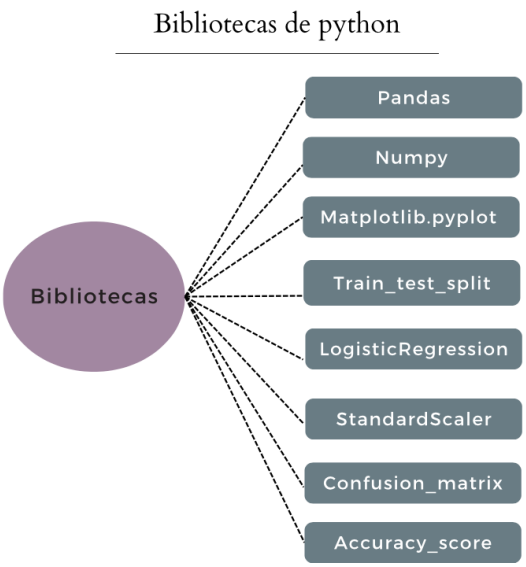

Figura 3 Bibliotecas

Posteriormente se eligen las variables que se utilizaran como características en X y la variable a predecir en la Y como se muestra en la Tabla 1.Este paso es la etapa de ingreso de datos, según la Figura 2.

Tabla 1: Características X

Edad	DebtRatio
45	0.80298213
40	0.1218762

La Tabla 2 se agrega en Y la variable dependiente a predecir como se muestra en la Tabla 2.

Tabla 2: Variable dependiente

SeriousDlqin2yrs
1
0

Continuando se divide la muestra test y entreno para poder entrenar el modelo y con el pseudocódigo de la Figura 4 la variable a predecir se entrena y da inicio a las primeras pruebas con los datos de test.

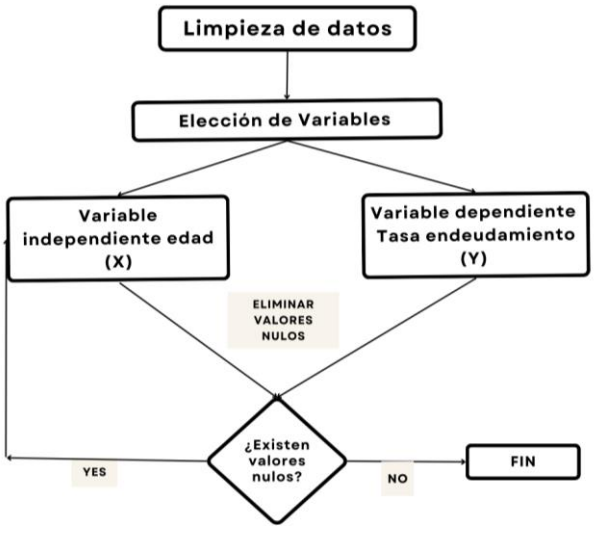

Figura 4 Pseudocodigo división valores(X,Y)

Después del entreno y definición de variables se saca la matriz de confusión y a su vez se valida en nivel de presición según la Figura 5.
Aquí se indica los valores falsos positivos y viseversa los negativos positivos.

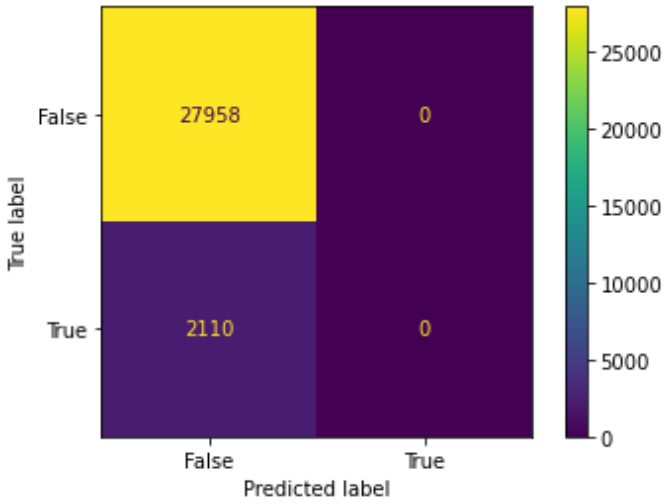

Figura 5 Matriz de confusion

Y como último paso se crea la gráfica como la Figura 6, de acuerdo con las variables de entrada que se definieron.
Se puede apreciar en la gráfica datos que se encuentran fuera de la muestra que son los 2 rojos y el verde.

Mientras tanto los demás se encuentran en el mismo rango la mayoría.

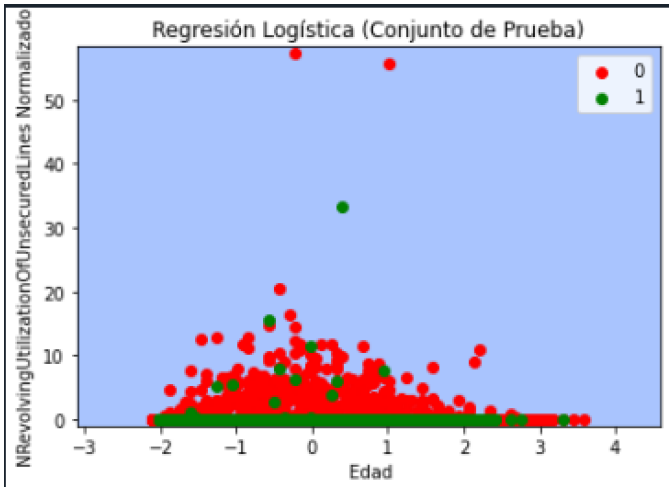

Figura 6 Gráfica

Para la predicción con árboles de decisión se sigue el mismo proceso hasta la declaración del modelo, como la Figura 7.

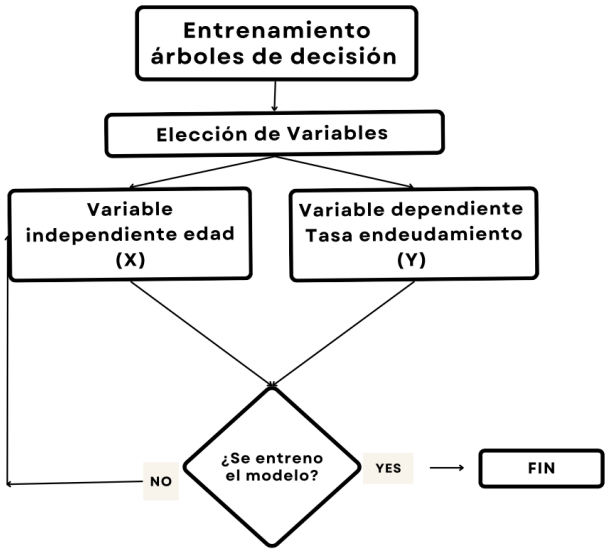

Figura 7 Árboles de decisión

Para la evaluación de estos dos modelos se pueden entrenar juntos y medir la precisión de cada uno según la Figura 8, es decir que se debe ser capaz de decirnos que modelo es el que tiene mayor precisión al momento de entrenar los datos.

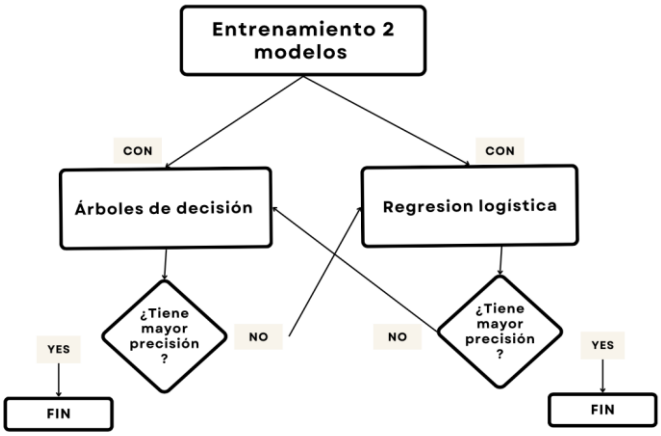

Figura 8 Comparación de modelos

Al estar evaluando constantemente es possible decir que modelo es el adecuado para poder validar un crédito bancario.

Esto mencionado se presenta en los resultados.

3. Resultados

Los resultados como se muestra en la Figura 9 se puede ver que la regresión logística tiene mayor precisión que los árboles, la diferencia es muy mínima, pero si se hablara de un volumen alto de datos, realmente el modelo más adecuado funcionaria mejor e incluso el tiempo de ejecución sería muy notorio.

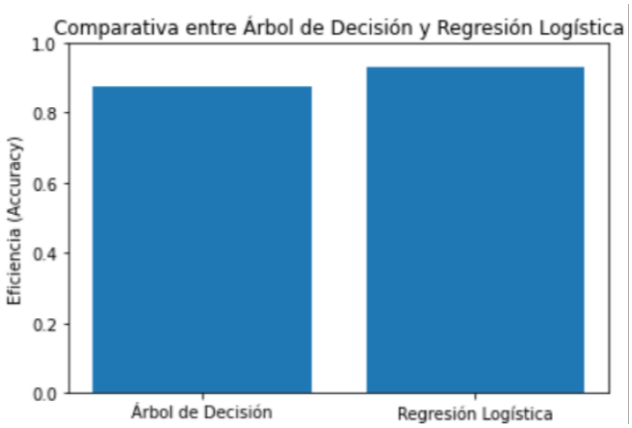

Figura 9 Gráfica resultados

Con este código se puede notar que la precisión de la regresión está por encima de los árboles de decisión como se muestra en la Figura 10.

```
Precisión promedio de Regresión Logística: 0.93
Precisión promedio de Árbol de Decisión: 0.87
```

Figura 10 Resultados

Con este resultado se nota que a pesar de que los árboles de decisión tienen ventajas, no siempre suele ser tan eficaz contra regresión logística.

Modelo	Precisión (%)
Árbol de decisión	0.87
Regresión logística	0.93

4. Discusión

La regresión logística superó al Árbol de decisión en términos de precisión, con una diferencia del 0.6%. Esto indica que la Regresión Logística logró clasificar correctamente a un mayor número de clientes en sus respectivas categorías de riesgo crediticio en comparación con el Árbol de Decisión.

5. Conclusiones

En conclusión, tanto los árboles de decisión como la regresión lineal son técnicas útiles en el análisis del riesgo crediticio, pero tienen diferencias significativas en su enfoque y aplicaciones. La regresión lineal se centra en modelar la relación lineal entre una variable dependiente (como el riesgo crediticio) y una o más variables independientes. Es especialmente útil cuando se busca entender la relación continua entre las variables y predecir el nivel de riesgo asociado con un prestatario en función de sus características específicas. La regresión lineal proporciona una medida cuantitativa del riesgo crediticio y permite analizar la importancia relativa de las diferentes variables en la predicción.

Por otro lado, los árboles de decisión se basan en una estructura de preguntas y respuestas que conducen a decisiones binarias. Estos árboles dividen los datos en ramas y nodos, y cada nodo

representa una pregunta o condición basada en una variable. Los árboles de decisión son especialmente útiles para identificar patrones y relaciones complejas entre las variables y clasificar a los prestatarios en diferentes categorías de riesgo. También son fáciles de interpretar y proporcionan una comprensión clara del razonamiento detrás de las decisiones tomadas.

6. Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de este estudio sobre la comparativa entre Árbol de Decisión y Regresión Logística para la predicción de riesgo crediticio. Su apoyo y contribución fueron fundamentales para el desarrollo de este trabajo.

En primer lugar, agradezco a profesor Francisco Jacob Ávila Camacho por proporcionar los datos lo cual fue esencial para llevar a cabo el análisis y las pruebas de los modelos. Su colaboración fue invaluable y sin ella este estudio no habría sido posible.

Referencias

Díaz, C. M., & Del Valle Guerra, Y. (2017). Riesgo financiero en los créditos al consumo del sistema bancario venezolano 2008-2015. *Orbis. Revista Científica Ciencias Humanas*, 20-40.

Diego Borrero-Tigreros, O. B.-L. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías*, √.

Fernández Castaño, H., & Pérez Ramírez, F. O. (2005). El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito. *Revista Ingenierías Universidad de Medellín, 4*(6), 55-75.

Hernández, P. A. (2004). Aplicacion de arboles de decision en modelos de riesgo crediticio. *Revista Colombiana de Estadıstica, 27*(2), 139 a 151.

Microsoft. (2016). *Cómo elegir algoritmos para Aprendizaje automático de Microsoft Azure.* Obtenido de https://docs.microsoft.com/es-es/azure/machine-learning/machine-learning-

MIRANDA, D. C. (13 de 06 de 2013). *EPISTEMUS*. Obtenido de www.epistemus.uson.mx: https://biblat.unam.mx/hevila/EpistemusCienciatecnologiaysalud/2013/no14/3.pdf

Moreno Villalba, L. M., Avila Camacho, F. J., & Melendez Ramirez, A. (4 de 1 de 2019). Intérprete de señales cerebrales con aprendizaje profundo para el control de Servomotores. *Tecnocultura, 17*(48), 12-16. Recuperado el 12 de julio de 2023, de https://tecnocultura.org/index.php/Tecnocultura/article/view/180

Pérez, J. (2017). La regresión logística como modelo de predicción del riesgo crediticio en las organizaciones de la economía social y solidaria. *Universidad Internacional del Ecuador (UIDE)*, 243.

Rodrigo, J. A. (10 de 2020). *Árboles de decisión con Python: regresión y clasificación*. Recuperado el 13 de 07 de 2023, de https://cienciadedatos.net/documentos/py07_arboles_decision_python.html

Vergara, J. G. (2022). Diseño de un modelo predictivo para otorgar créditos. *SICIELO*, 320-347.